PREDICTING LOAN REPAYMENT

OVERVIEW

01 PREFACE

- Business Problem
- Data description

02 METHODOLOGY

- Data Pre-processing
- Exploratory Data Analysis

03 DATA VISULIZATION

- Data Interpretation
- Business Insights

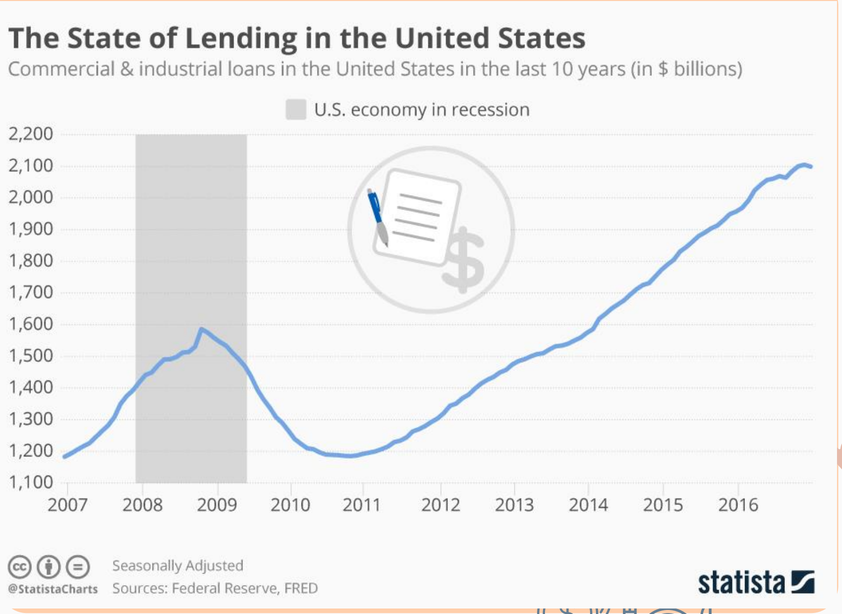
04 MODELING

- Logistic Regression
- Decision Tree
- Random Forest Classifier

01.
PREFACE

BUSINESS PROBLEM

- Financial institutions are facing the challenge of optimizing loan approval processes to strike a balance between expanding their customer base and mitigating the risk of non-performing loans.
- Prominent microfinance institutions operating in emerging markets reveal that up to 25% of their annual revenue is directly linked to the successful repayment of small-business loans.
- The objective is to enhance profitability by increasing the approval rate for credit applications while minimizing the likelihood of defaults, thereby ensuring sustainable financial performance and maintaining a healthy loan portfolio.



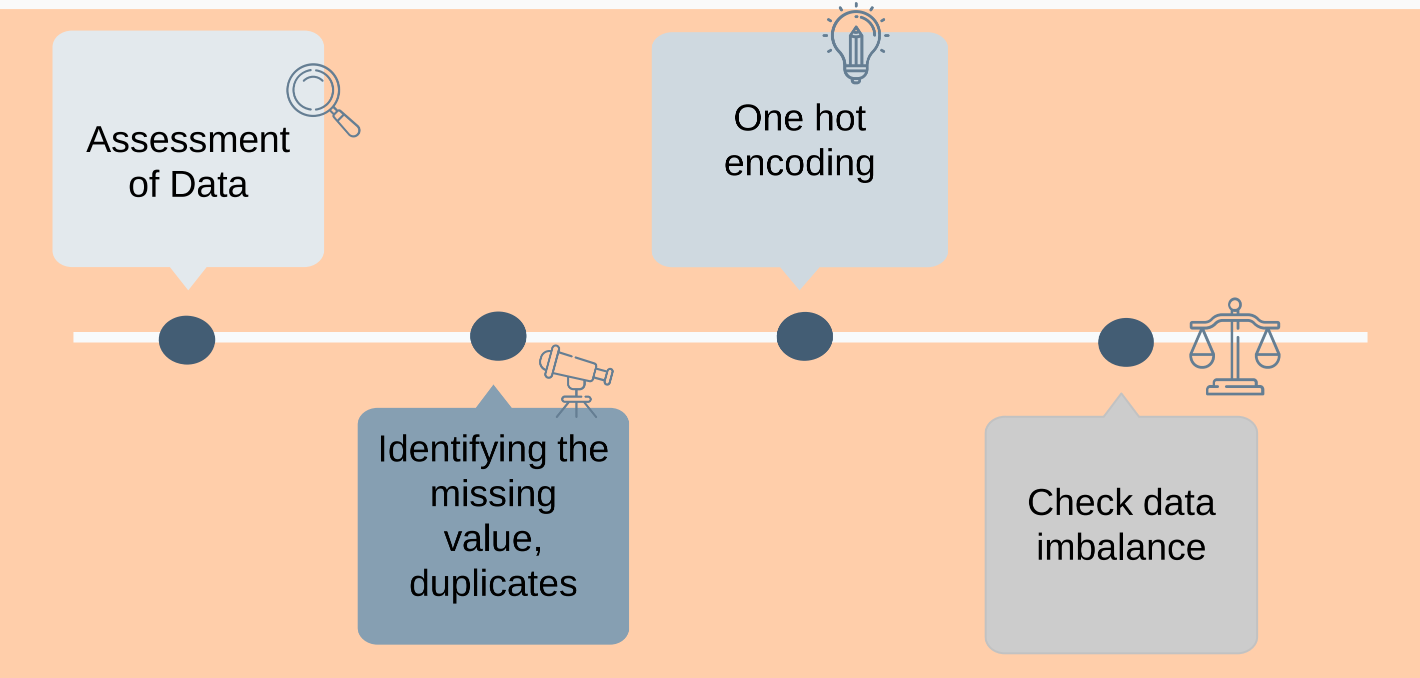
DATA DESCRIPTION

- Dataset Source https://www.kaggle.com/datasets/sarahvch/predicting-who-pays-back-loans/data
- 10k rows
- 14 Features

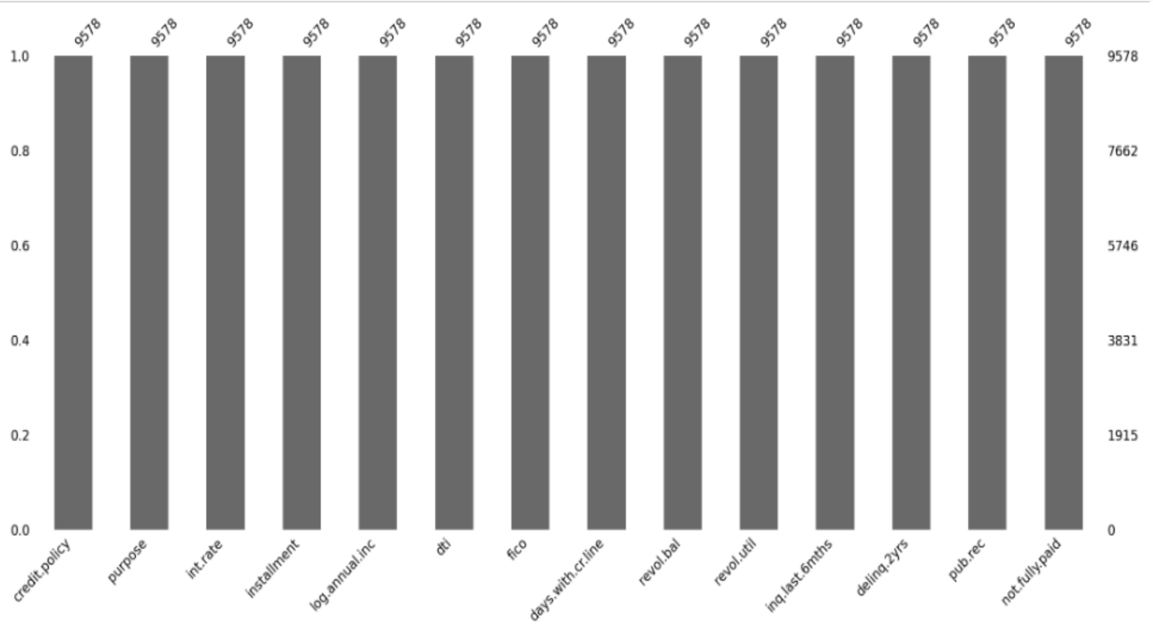
credit policy	1 if the customer meets the credit underwriting criteria of LendingClub.com, and 0 otherwise
purpose	The purpose of the loan like credit card, small business, etc
int rate	The interest rate of the loan (proportion)
installment	The monthly installments owed by borrower if loan is funded
log annual inc	The natural log of the annual income of borrower
dti	The debt-to-income ratio of the borrower
fico	The FICO credit score of the borrower.
days with cr line	The number of days the borrower has had credit line
revol bal	The borrower's revolving balance
revol util	The borrower's revolving line utilization rate
inq last 6mths	The borrower's number of inquiries by creditors in the last 6 months
delinq 2yrs	The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
pub rec	The borrower's number of derogatory public records
not fully paid	indicates whether the loan was not paid back in full

02.
METHODOLOGY

EXPLORATORY DATA ANALYSIS



ASSESSMENT OF DATA



Plotting the missing values using the missingno package

Inference:
There is no null values

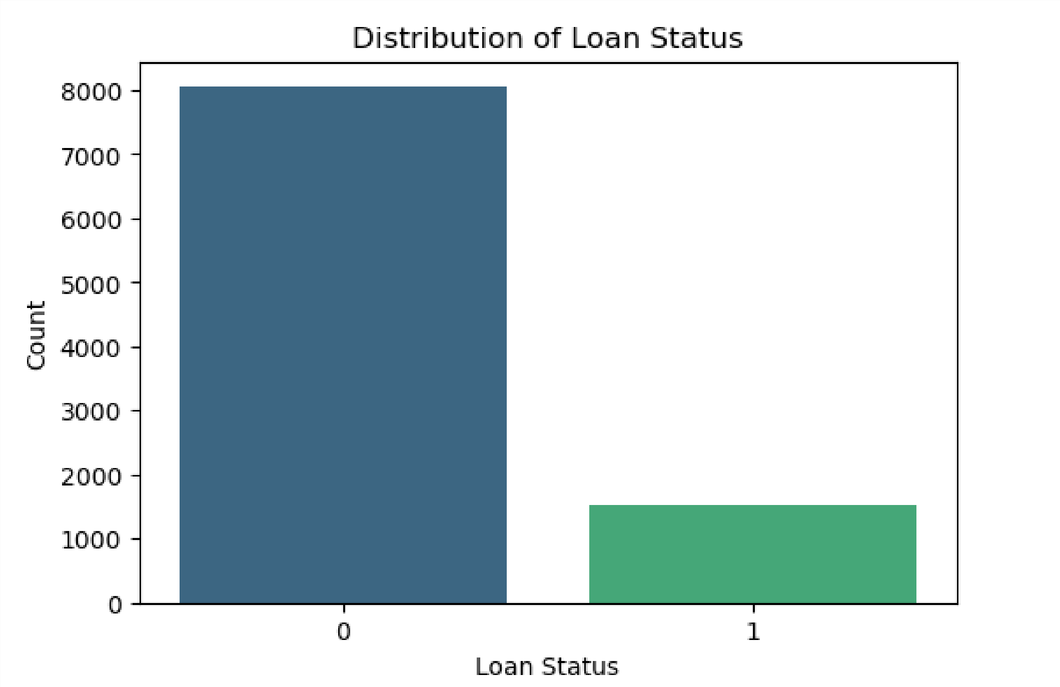
ONE HOT ENCODING

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9578 entries, 0 to 9577
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   credit.policy      9578 non-null   int64
 1   purpose            9578 non-null   object
 2   int.rate           9578 non-null   float64
 3   installment        9578 non-null   float64
 4   log.annual.inc     9578 non-null   float64
 5   dti                9578 non-null   float64
 6   fico               9578 non-null   int64
 7   days.with.cr.line  9578 non-null   float64
 8   revol.bal          9578 non-null   int64
 9   revol.util         9578 non-null   float64
 10  inq.last.6mths     9578 non-null   int64
 11  delinq.2yrs        9578 non-null   int64
 12  pub.rec            9578 non-null   int64
 13  not.fully.paid     9578 non-null   int64
dtypes: float64(6), int64(7), object(1)
memory usage: 1.0+ MB
```

One hot encoding transforms categorical data into numerical - it transforms strings into numbers so that we can apply our Machine Learning algorithms without any problems.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9578 entries, 0 to 9577
Data columns (total 19 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   credit.policy       9578 non-null   int64
 1   int.rate            9578 non-null   float64
 2   installment         9578 non-null   float64
 3   log.annual.inc      9578 non-null   float64
 4   dti                 9578 non-null   float64
 5   fico                9578 non-null   int64
 6   days.with.cr.line   9578 non-null   float64
 7   revol.bal           9578 non-null   int64
 8   revol.util          9578 non-null   float64
 9   inq.last.6mths      9578 non-null   int64
 10  delinq.2yrs         9578 non-null   int64
 11  pub.rec             9578 non-null   int64
 12  not.fully.paid      9578 non-null   int64
 13  credit_card         9578 non-null   uint8
 14  debt_consolidation  9578 non-null   uint8
 15  educational         9578 non-null   uint8
 16  home_improvement    9578 non-null   uint8
 17  major_purchase      9578 non-null   uint8
 18  small_business      9578 non-null   uint8
dtypes: float64(6), int64(7), uint8(6)
memory usage: 1.0 MB
```

CHECKING DATA IMBALANCE



The data is very imbalanced. So we're going to use bagging techniques.

03.
DATA
VISUALIZATION

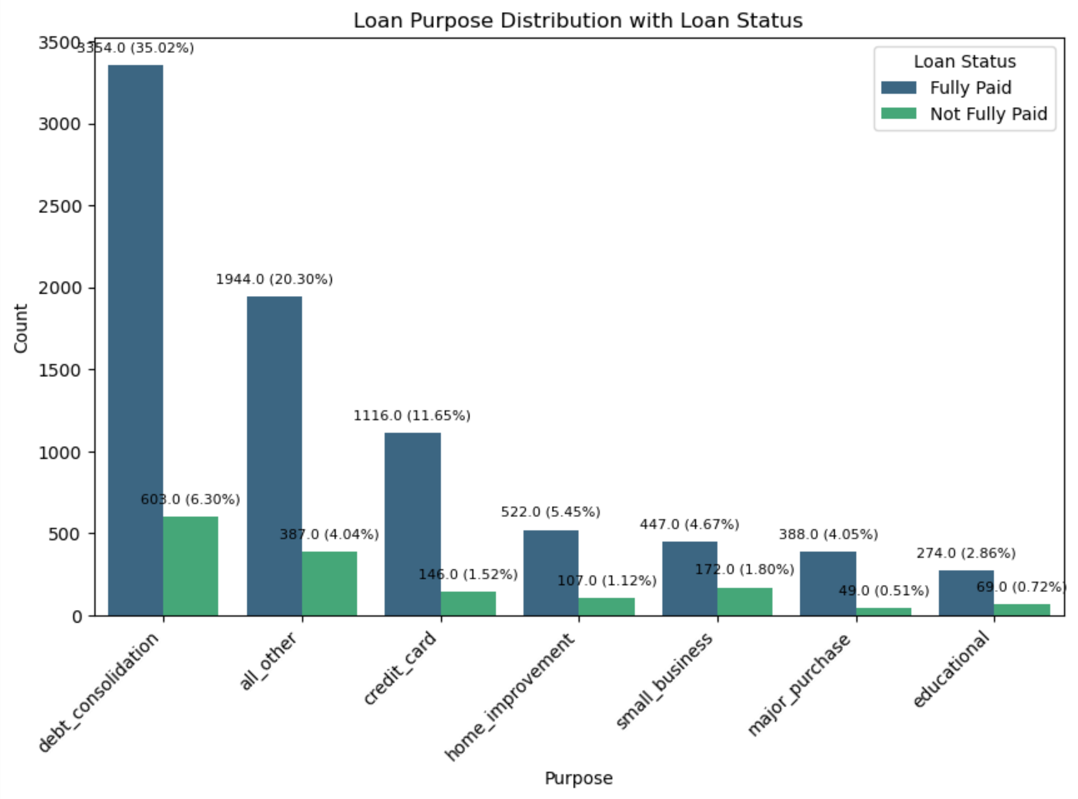
DATA VISUALIZATION

- This graph shows the distribution of Loan purpose vs Loan Status.

- The distribution shows majority of the loan attributes to debt_consolidation purpose

- Credit_card, Major_purchase & debt_consolidation has higher repayment ratio compared to categories.

- Small business and home repayment has least the repayment ratio.

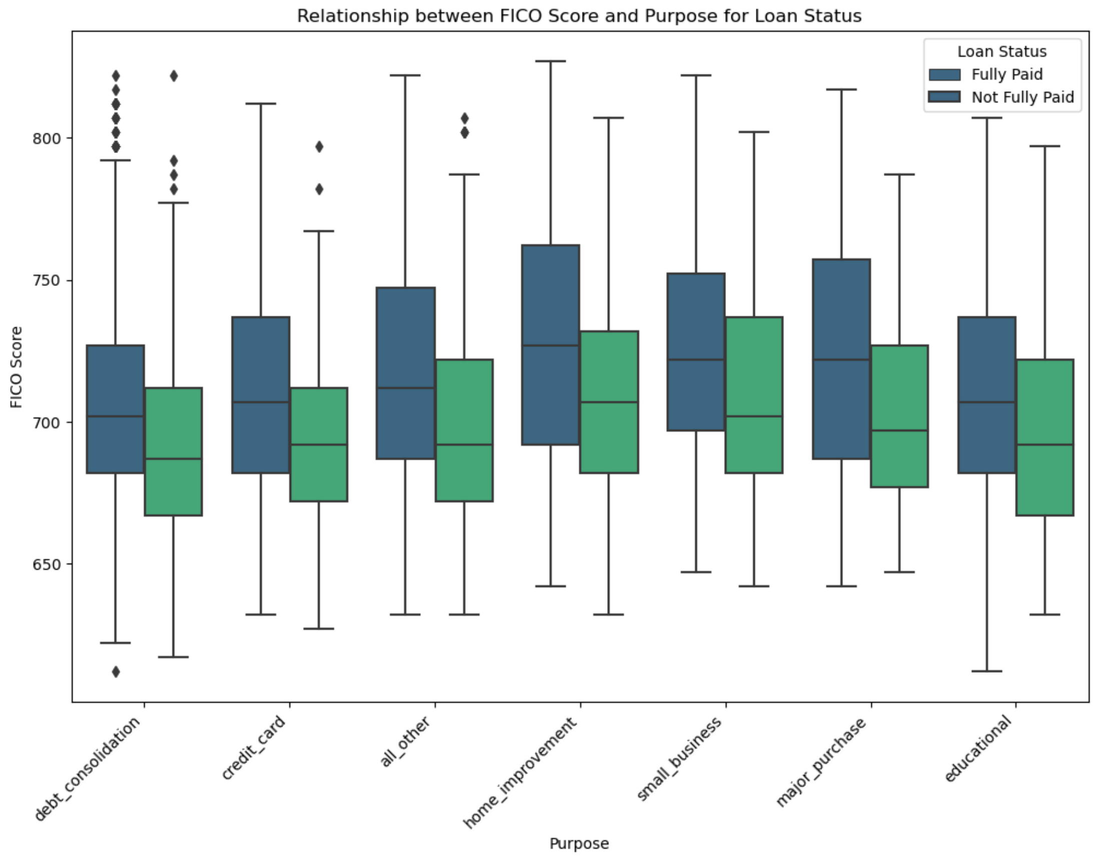


DATA VISUALIZATION

- This box-plot shows the comparison of FICO w.r.t Loan Status across each category.

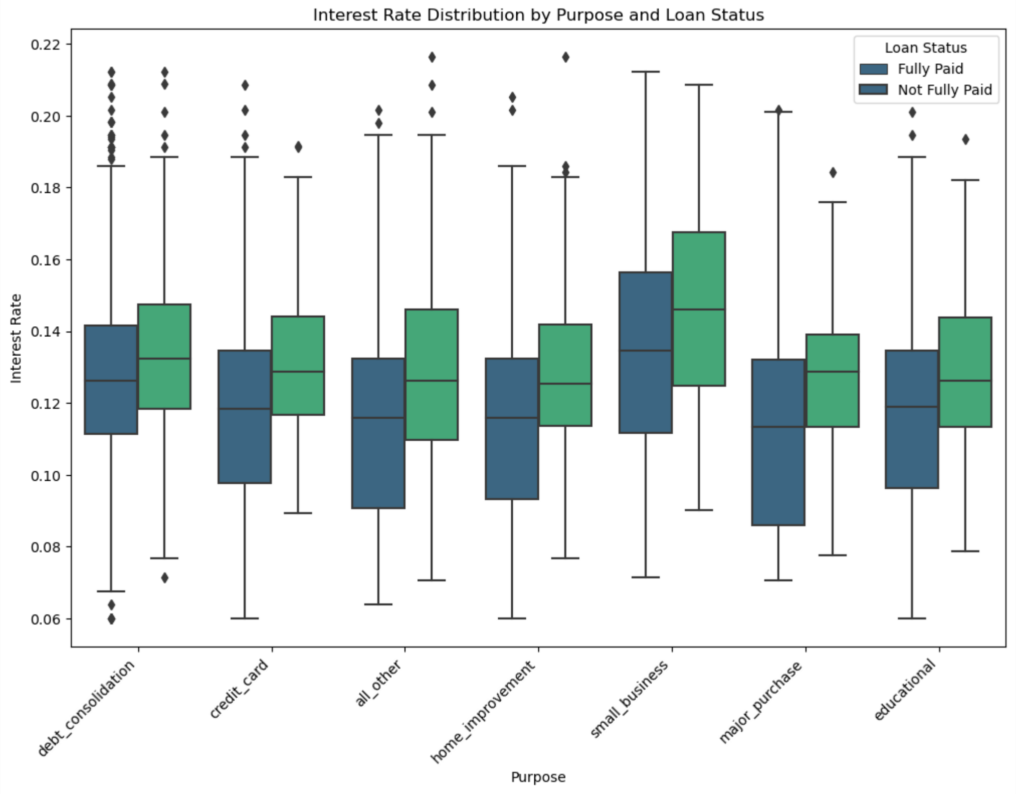
- It can be inferred that FICO score is relatively higher in comparison to Fully paid vs Not Fully paid.

- The median of the Box plot of Not Fully paid is Right skewed which means majority do have lower FICO score



DATA VISUALIZATION

- This graph shows the interest rate distribution of the loan for the various purpose
- Small_business category has higher median interest rate and major_purchase has lower median interest rate.
- Loans of Not Fully paid has increased median interest rate compared to Fully paid one's.

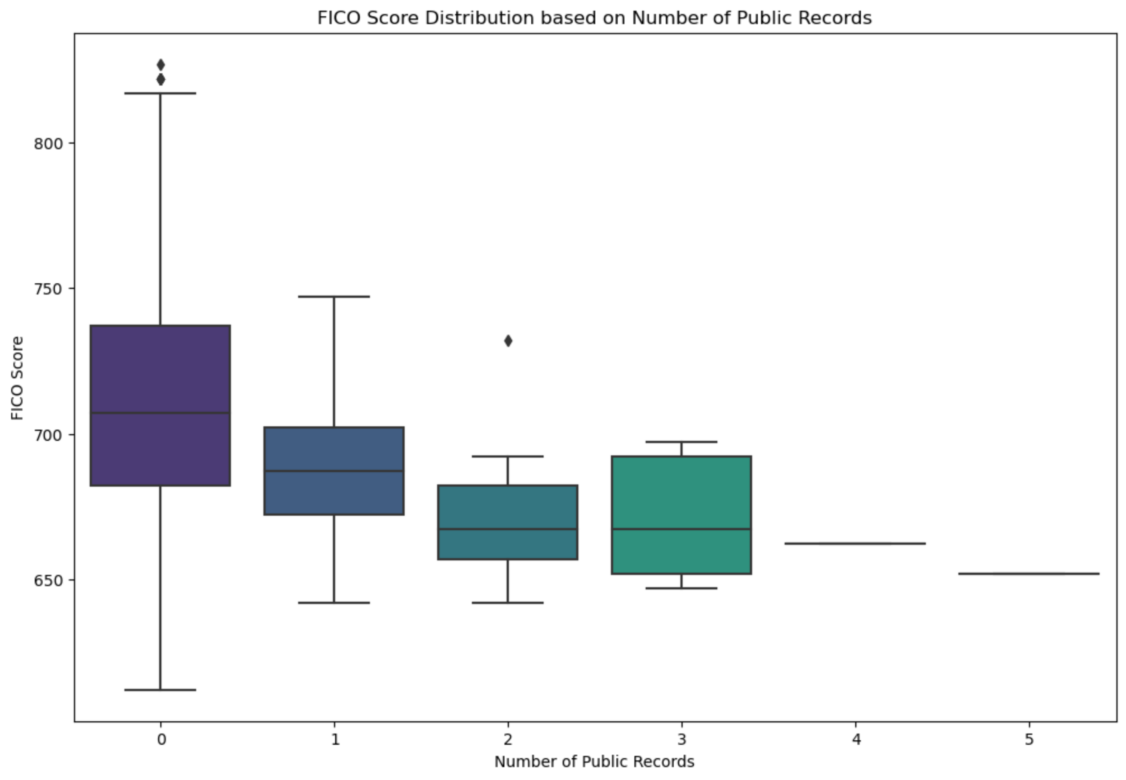


DATA VISUALIZATION

- This graph shows the FICO comparison with w.r.t no of public records.

- Majority of the borrowers with higher score have least no of records.

- FICO score is inversely correlated with No of public records



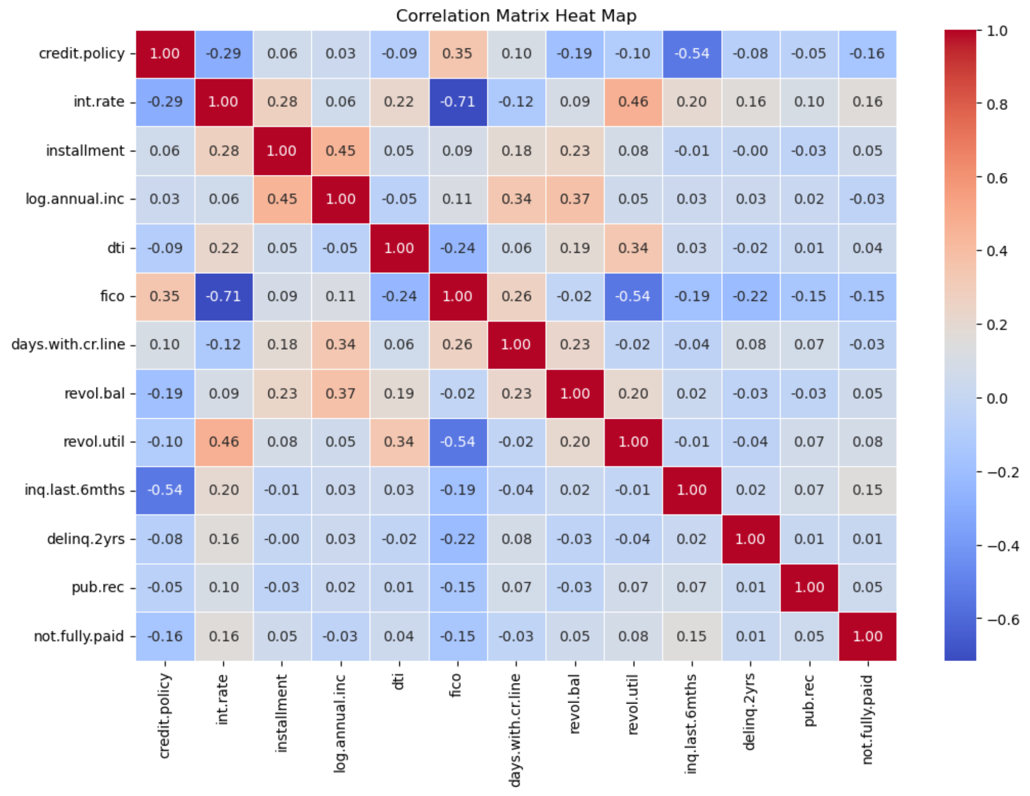
DATA VISUALIZATION

- The Heat map shows the correlation across each category variables

- Interest rate and revol.util rate are moderatly correlated.

- Also, Credit policy and FICO score are somehow correlated

- Thus, Random forest model would be used as baseline model to better understand feature and immune to multicollinearity



04.
MODELING

LOGISTIC REGRESSION

```
[[2403    5]
 [ 459    7]]

Recall:  0.015021459227467811
Precision:  0.5833333333333334
F-1 score:  0.029288702928870296
Accuracy:  0.8385525400139179
```

```
[[1477  931]
 [ 212  254]]

Recall:  0.5450643776824035
Precision:  0.21434599156118145
F-1 score:  0.3076923076923077
Accuracy:  0.6022964509394572
```

Logistic Regression:

- The recall is very low, indicating poor performance in capturing positive instances.
- The precision is relatively high, suggesting that when the model predicts positive, it's correct.
- The overall accuracy is high, but it might be misleading due to the imbalanced nature of the data.

Logistic Regression with Balanced Bagging:

- The recall has improved significantly compared to the regular Logistic Regression, indicating better performance in capturing positive instances.
- The precision is lower than in regular Logistic Regression.
- Accuracy decreased, but the performance is stable

DECISION TREE using BALANCED BAGGING

```
[[1502  906]
 [ 166  300]]

Recall:  0.6437768240343348
Precision is:  0.24875621890547264
F-1 score is:  0.358851674641483
Accuracy:  0.627000695894224
```

- It can be observed that the accuracy increases from 0.6 as per logistic regression to 0.627 as per decision tree using balanced bagging
- As we are focused more on recall, we observe that it increases from 0.54 to 0.64, which indicates improved positive instance predictions
- Precision also increases from 0.21 to 0.24

BALANCED RANDOM FOREST CLASSIFIER

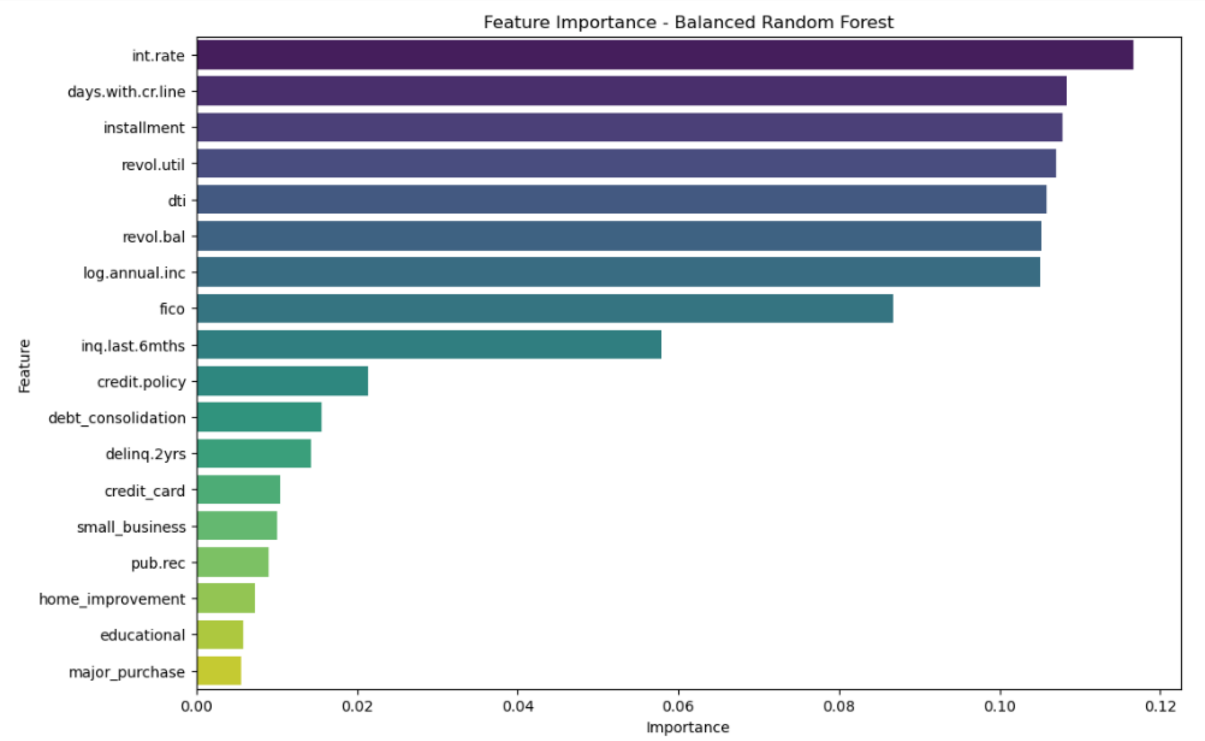
```
[[1423  985]
 [ 162  304]]

Recall:  0.6523605150214592
Precision is:  0.2358417377812257 5
F-1 score is:  0.3464387464387464
Accuracy:  0.6009046624913014
```

- As we are focused more on **recall,** we observe that it increases from 0.64 as per decision tree model to 0.65 as per balanced random forest classifier indicating slight increase in positive predictions

- Specificity decreases from 0.35 as per decision tree to 0.34 as per random forest classifier

- Accuracy also decreased from 0.62 to 0.60

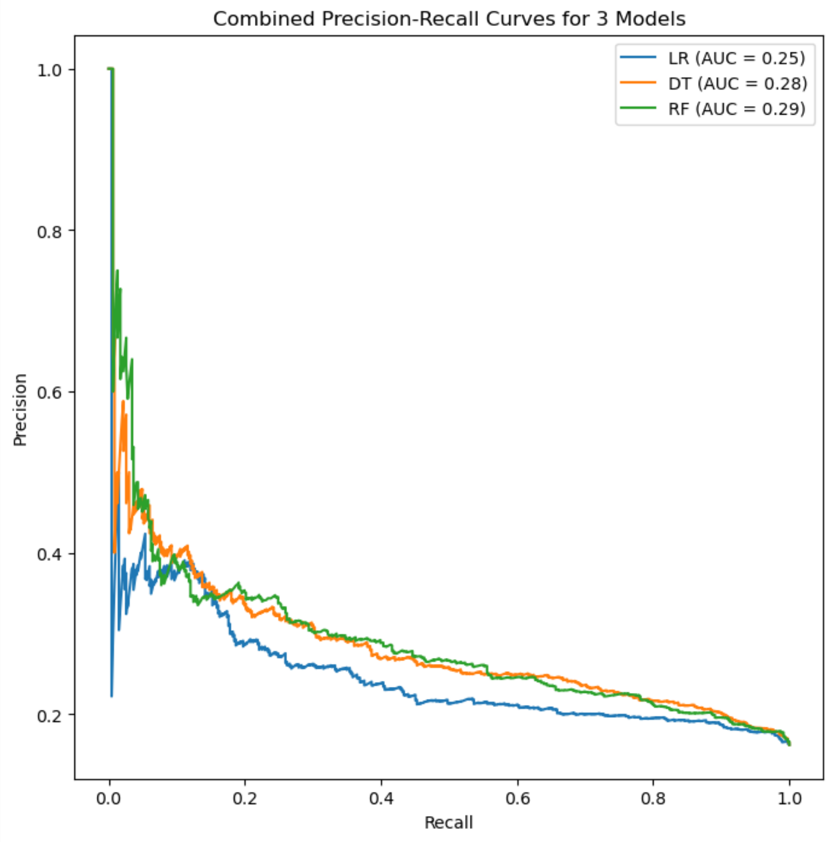
BALANCED RANDOM FOREST CLASSIFIER



- Random Forest Graph depicts the most important variables used by the model.

- The order of importance from top to bottom

MODEL COMPARISON



- In examining Precision-Recall curves, it's crucial to highlight that the Random Forest (RF) model stands out with the highest Area Under the Curve (AUC) value at 0.29.

- This signifies its superior ability to capture positive instances, a key strength aligned with our dataset's primary goal.

CONCLUSION

- Performed Logistic regression, Decision Tree and Balanced Random Forest Classifier models
- Chose the model based on highest recall i.e True Positives as the positive class is 1, which signifies that the model should be able to predict higher defaulters.
- Random Forest classifier has the Best Recall amongst all models
- To tackle data imbalance, employ techniques such as oversampling and undersampling.
- Utilize feature engineering like scaling and binning for enhanced models.
- Explore advanced algorithms such as SVM, Gradient Boosting Machines, and Neural Networks.

Thank You